

Self-organization in social tagging systems

Chuang Liu,^{1,2} Chi Ho Yeung,^{3,*} and Zi-Ke Zhang^{3,4,†}

¹*School of Business, East China University of Science and Technology, Shanghai 200237, China*

²*Engineering Research Center of Process Systems Engineering (Ministry of Education),
East China University of Science and Technology, Shanghai 200237, China*

³*Department of Physics, University of Fribourg, Fribourg CH-1700, Switzerland*

⁴*Web Sciences Center, University of Electronic Science and Technology of China - Chengdu 610054, PRC*

(Dated: January 12, 2013)

Individuals often imitate each other to fall into the typical group, leading to a self-organized state of typical behaviors in a community. In this paper, we model self-organization in social tagging systems and illustrate the underlying interaction and dynamics. Specifically, we introduce a model in which individuals adjust their own tagging tendency to imitate the average tagging tendency. We found that when users are of low confidence, they tend to imitate others and lead to a self-organized state with active tagging. On the other hand, when users are of high confidence and are stubborn for changes, tagging becomes inactive. We observe a phase transition at a critical level of user confidence when the system changes from one regime to the other. The distributions of post length obtained from the model are compared to real data which show good agreements.

PACS numbers: 89.65.-s, 89.20.Hh, 05.65.+b

I. INTRODUCTION

Self-organization is an interesting phenomenon observed in various areas including network growth [1], traffic jams [2] and resource allocation [3]. In social systems, individuals often imitate each other through interaction and observation, to become more typical in the community. Such dynamics results in a steady state in which most individuals adopt the typical practice by learning from each other. In online communities, self-organization is further facilitated by the recent advent of Web 2.0 social applications, which encourage Internet users to interact with peers. By interacting with each other, users self-organize and lead to a state of typical behaviors.

In resource sharing applications, tags are practical to facilitate the search and management of resources [4, 5]. Tags are usually simple labels and annotations which help users to have preliminary understanding of the content before collecting the resources. Recently, *tagging systems* are implemented in popular applications including *delicious.com*, *flickr.com* and *citeulike.org*. To well organize their resources, users assign tags with their bookmark, pictures or Bibtex files. By browsing through tags, users are able to find other users who share similar interests. Tags thus reflect user behaviors and preferences, and with which ones can easily search, collaborate and form communities with others [6].

Tagging systems are studied extensively in recent years, but the underlying interaction and dynamics among tag users are still unclear. Mathematically, tagging systems are composed of fundamental units of user-resource-tag triples [5, 7, 8], and each tagging action con-

stitutes one or several hyper-links in a tripartite graph. Such user-resource-tag relations are often referred to as *folksonomy*. Examples include the use of keywords or PACS numbers in academic papers, which also helps to reveal the structure of citation networks [9, 10]. However, how similar papers influence each other on the choice of keywords is still an open question. To reveal the tagging dynamics, Cattuto *et al* [11] suggested to consider the process of social annotation as a collective yet uncoordinated exploration of the underlying semantic space through a series of random walks. In Ref. [12], Lambiotte *et al* modeled folksonomy in terms of tripartite graphs. Zhang and Liu [13] proposed a model to explain some statistical properties in folksonomy, in which users can search for resources via tags. Many of these studies consider individual tag assignment, while ignoring the interaction among peer tag users.

In this paper, we propose a model to investigate the dynamics and interaction among individuals in a tagging system. Specifically, individuals imitate each other in tagging which results in a self-organized state. We found that when users are of low confidence, they self-organize to attain a steady state of active tagging. On the other hand, the system ends with inactive tagging when users are confident of their own tagging practice. In addition, a phase transition is observed with a critical level of user confidence, when the system changes from one regime to the other. Furthermore, we compare distributions of post length from the proposed model to two real datasets obtained from *delicious.com* and *flickr.com*, which show good agreements.

II. MODEL

We consider a model of tagging system with N users. At each step, each user posts one resource and assign

*chbyeung@gmail.com

†zhangzike@gmail.com

tags to the resource. The tendency of which the user assigns tags is characterized by $p_i(t)$, which is the probability that the user continues with tag assignment for the resource. In other words, the probability that user i assigns $n_i(t)$ tags at time t is given by

$$\Pr[n_i(t) = l] = p_i^l(t)[1 - p_i(t)], \quad (1)$$

where $l = 1, 2, 3, \dots$. Large $p_i(t)$ corresponds to a high tendency to assign tags and vice versa. We thus call $p_i(t)$ the *tagivity*, which characterizes the tendency of user i in tag assignment. Given that $p_i(t)$ remains unchanged, $n_i(t)$ follows a geometric distribution with parameter $1 - p_i(t)$. We model the self-organization of user by assuming that users adjust their $p_i(t)$ based on the observation of $\langle p(t) \rangle$, the average tagivity over all users at time t .

As one main purpose for tagging is to facilitate the search of resources for others, users would tend to adopt a more typical tagging practice. They thus adjust their own tagivity in order to imitate the observed average tagivity over users. We denote the combination of tags associated with a resource to be a *post*. Based on observations, users obtain an estimated distribution of *post length*, which is the number of tags associated with each post. We assume that the users estimate the distribution based on the average user tagivity, as given by

$$\Pr[l' = l] = \langle p(t) \rangle^l [1 - \langle p(t) \rangle], \quad (2)$$

where l' corresponds to the observed post length. With this distribution in mind, user i randomly picks a post and imitates its length in the next step. Suppose user i assigns $n_i(t)$ tags at time t , the probability that he/she picks a post of length l' less than $n_i(t)$ is given by

$$\Pr[l' < n_i(t)] = 1 - \langle p(t) \rangle^{n_i(t)-1}. \quad (3)$$

On the other hand, the probability that user i picks a post of length l' larger than $n_i(t)$ is given by

$$\Pr[l' > n_i(t)] = \langle p(t) \rangle^{n_i(t)}. \quad (4)$$

With probability $\langle p(t) \rangle^{n_i(t)-1} (1 - \langle p(t) \rangle)$, user i picks a post of length equals to his/her own post length at time t .

Users imitate the post they pick up by changing their tagivities. For instance, user i increases his/her tagivities if $n_i(t)$ is smaller than l' , and vice versa. We denote the probabilities of which user i increases, maintains or decreases his/her tagivity as $\eta_i^+(t)$, $\eta_i^0(t)$ and $\eta_i^-(t)$, given by

$$\eta_i^+(t) = \frac{(1 - \beta) \langle p(t) \rangle^{n_i(t)}}{Z_i(t)}, \quad (5)$$

$$\eta_i^0(t) = \frac{\beta [\langle p(t) \rangle^{n_i(t)-1} (1 - \langle p(t) \rangle)]}{Z_i(t)}, \quad (6)$$

$$\eta_i^-(t) = \frac{(1 - \beta) (1 - \langle p(t) \rangle^{n_i(t)-1})}{Z_i(t)}, \quad (7)$$

where $Z_i(t)$ ensures $\eta_i^+(t) + \eta_i^0(t) + \eta_i^-(t) = 1$. The parameter $\beta \in [0, 1]$ can be considered as the *confidence* of user on his own tagivity: $\beta = 0$ corresponds to the case with *unconfident* users who tend to change their choice of tagivities every time step, and $\beta = 1$ corresponds to the case with *confident* users who stay with their tagivities every time step. Increasing β from 0 to 1 characterizes the increase in user confidence, such that users are more reluctant to changes.

We propose two response functions based on which the tagivity is updated. In the first case, the tagivity is updated *linearly* by

$$p_i(t+1) = p_i(t) + a_i(t) \delta_l, \quad (8)$$

where $a_i(t) = 1, 0, -1$ respectively with probabilities $\eta_i^+(t)$, $\eta_i^0(t)$, $\eta_i^-(t)$, and $\delta_l > 0$ is a parameter which characterizes the extent the tagivity is changed. When $a_i(t) = 1$ or -1 , the tagivity increases or decreases. The parameter δ_l can be interpreted as the *adaptability* of the users. Large δ_l corresponds to faster adaptation to the typical behaviors.

In the second case, the complementary tagivity $1 - p_i(t)$ is updated *multiplicatively* by

$$1 - p_i(t+1) = [1 - p_i(t)] (1 + \delta_m)^{-a_i(t)}, \quad (9)$$

where $\delta_m \geq 0$ serves the same role as δ_l in linear update. A more explicit implication of this multiplicative updating can be obtained by the relation $E[n_i(t)] = (1 - p_i(t))^{-1}$, where $E[n_i(t)]$ is the expected value of $n_i(t)$ based on the geometric distribution. Equation (9) thus implies

$$E[n_i(t+1)] = E[n_i(t)] (1 + \delta_m)^{a_i(t)}. \quad (10)$$

In other words, the expected value of $n_i(t)$ respectively increases by a factor of $(1 + \delta_m)$, remains unchanged or decreases by a factor of $(1 + \delta_m)^{-1}$ with $a_i(t) = 1, 0 - 1$.

III. SIMULATION RESULTS

To reveal the dynamics underlying self-organization in the model, we conduct numerical simulations. We start with random initial $p_i(0)$ for all users. At time t , $n_i(t)$ is drawn according to the probabilities in Eq. (1), such that η_i^+ , η_i^0 and η_i^- are evaluated according to Eqs. (5)-(7). The tagivity $p_i(t)$ for each user is then updated according to Eq. (8) in the case of linear update or Eq. (9) in the case of multiplicative update. Unless specified, the results are obtained when the system converges, i.e. $\langle p(t) \rangle$ becomes steady. We observed that $\langle p(t) \rangle$ has a slight fluctuation around a time average value and the fluctuation is dependent on δ_l and δ_m .

A. Convergence time

We first study the relation between the convergence time and the parameters β , δ_l and δ_m . The self-organized

state in our context corresponds to the state in which $\langle p(t) \rangle$ becomes steady. The convergence time τ is thus defined by the relation $\langle p(\tau) \rangle \approx \langle p(\tau + L) \rangle$, for all $t \geq \tau$ and some sufficiently large L .

The convergence time is plotted in Fig. 1(a) as a function of confidence β . As similar results are obtained from the two update rules, we present only the results obtained from the linear update. As shown in Fig. 1(a), the larger the adaptability δ_l , the faster the convergence time. The prominent peaks of convergence time observed at $\beta \approx 0.5$ suggest the possibility of a phase transition at $\beta \approx 0.5$ as dynamics slows down. Furthermore, peak positions are similar at different values of δ_l . It implies that, when the weight β in Eqs. (5) - (7) to modify tagivity is equal to that to maintain tagivity, the users are confused and the self-organization slows down. As the convergence time is also dependent on system size, we plot in log-log scale N as a function of τ at $\beta = 0.5$ in Fig. 1(b), as studies [14, 15] suggest a conventional scaling of $\ln \tau \propto \ln N$ in the proximity of phase transition. These results suggest that on top of the self-organization, there is a phase transition in the range close to $\beta = 0.5$.

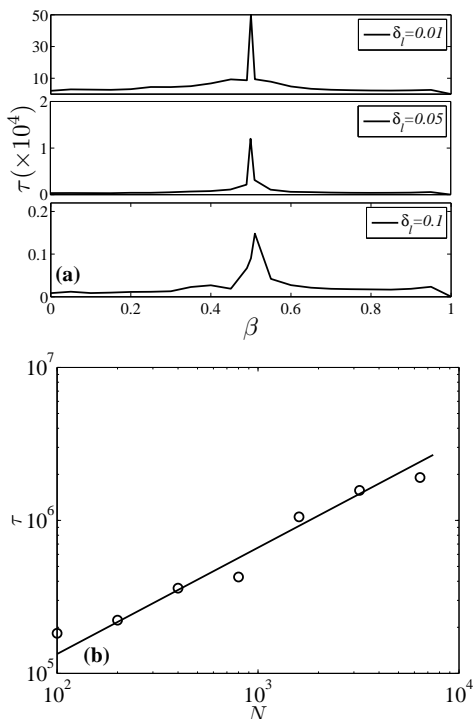


FIG. 1: (a) Convergence time τ as a function of β for different δ_l . Convergence time peaks around $\beta = 0.5$. The larger the adaptability δ_l , the more quickly the system reaches steady state. (b) Convergence time as a function of N when $\beta = 0.5$, which show scattering of data round the straight line implying $\ln \tau \propto \ln N$.

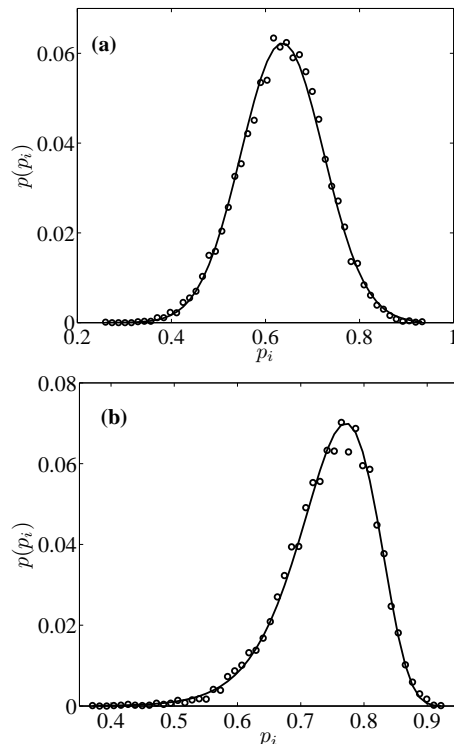


FIG. 2: The tagivity distributions with (a) the linear update and (b) the multiplicative update. Parameters: $\beta = 0.45$ and $\delta_l = 0.05$ for linear update and $\beta = 0.4$ and $\delta_m = 0.1$ for multiplicative update. Fittings: (a) Gaussian fit with $\mu = 0.63$ and $\sigma = 0.088$ and (b) log-normal fitting with $\mu = -1.41$ and $\sigma = 0.27$.

B. Steady distributions of tagivity

As mentioned in Sec. II, each user at each step randomly picks a post and imitates its length, their tagivities thus fluctuate around the average values. We show in Fig. 2 the stable distribution of tagivity after the system converges. Figure 2(a) shows that the stable distribution from linear update resembles Gaussian distribution. The simulation results are obtained by $\beta = 0.45$ and $\delta_l = 0.05$ and the parameter of Gaussian fit are $\mu = 0.63$ and $\sigma = 0.088$. The results are not as obvious as Eq. (8) suggests in the case when $a_i(t)$ is a random variable. In such case, $\sum_{t=1}^{\infty} a_i(t)$ would result in an infinite variance of $p_i(t)$, as compared to the finite variance observed in Fig. 2(a). The finite variance of $p_i(t)$ comes from the restoring process of $a_i(t)$ around the typical behaviors, as given by the probabilities in Eqs. (5) - (7). Figure 2(b) shows the stable distribution of tagivity obtained from the multiplicative update, where $1 - p_i$ approximately follows the log-normal distribution. Simulation results are obtained by $\beta = 0.4$ and $\delta_m = 0.1$, with log-normal fitting of $\mu = -1.41$ and $\sigma = 0.27$. The origin of the log-normal distribution is similar to that of Gaussian distribution and can be seen by taking algorithm of Eq. (9).

IV. PHASE TRANSITION AND SELF-ORGANIZATION

Though analytic solutions for the general case are difficult to obtain, we can write down a simple description of the steady state when $\delta_l \rightarrow 0$ or $\delta_m \rightarrow 0$. In this case we assume $p_i \approx \langle p \rangle$ for all user i . We further introduce a quantity Δ which characterizes the tendency for $\langle p \rangle$ to increase or decrease, as given by

$$\Delta(\langle p \rangle) = \sum_{n=1}^{\infty} \langle p \rangle^{n-1} (1 - \langle p \rangle) [\eta^+(n, \beta) - \eta^-(n, \beta)], \quad (11)$$

Δ describes the difference between η^+ and η^- when the average user tagivity is $\langle p \rangle$. A positive Δ corresponds to a tendency for $\langle p \rangle$ to increase, and vice versa. Substitutions of Eqs. (5) and (7) for η^+ and η^- into Eq. (11) lead to the following expression

$$\Delta(\langle p \rangle) = \sum_{n=1}^{\infty} \frac{\langle p \rangle^{n-1} (1 - \langle p \rangle) (\langle p \rangle^n - 1 + \langle p \rangle^{n-1})}{Z(n, \beta)}. \quad (12)$$

We numerically evaluate the summation in Eq. (12) and obtain the values of $\langle p \rangle$ when $\Delta = 0$, i.e. when there is no tendency for $\langle p \rangle$ to increase or decrease and the system becomes steady.

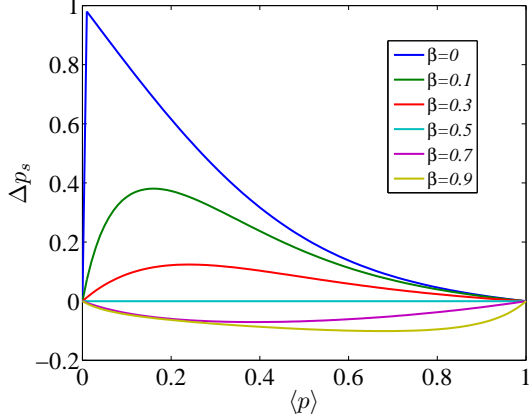


FIG. 3: (Color online) Δ as a function of $\langle p \rangle$ for different values of β .

Figure 3 shows Δ as a function of $\langle p \rangle$ for different β . These results imply that for all β , $\langle p \rangle = 0, 1$ are solutions of $\Delta = 0$. The fixed points of $\langle p \rangle = 0$ or $\langle p \rangle = 1$ respectively correspond to the cases when all users in the system stop active tagging or assign infinite number of tags. When $\beta < 0.5$, we get $\Delta \geq 0$ for all $\langle p \rangle$, which implies that the tendency to increase tagivity is higher than that to decrease tagivity, leading to a stable fixed point at $\langle p \rangle = 1$. On the other hand, when $\beta > 0.5$, we get $\Delta \leq 0$ which implies that the tendency for the tagivity to decrease is larger than that to increase, leading to an opposite result of stable fixed point at $\langle p \rangle = 0$. This drastic change of the self-organized state corresponds to

a phase transition at $\beta \approx 0.5$ from a regime with active tag assignment to one with inactive tag assignment. It is also interesting to note that when $\beta = 0.5$, $Z \equiv 1$ for all n in Eq. (12) such that $\Delta \equiv 0$ is guaranteed by the identity

$$\sum_{n=1}^{\infty} \langle p \rangle^{n-1} (1 - \langle p \rangle)^{n-1} - \langle p \rangle^n \equiv 0 \quad (13)$$

for all values of $\langle p \rangle$. It implies that at the critical point of $\beta = 0.5$, the system does not have a unique fixed point of $\langle p \rangle$, unlike the cases with $\beta \neq 0.5$.

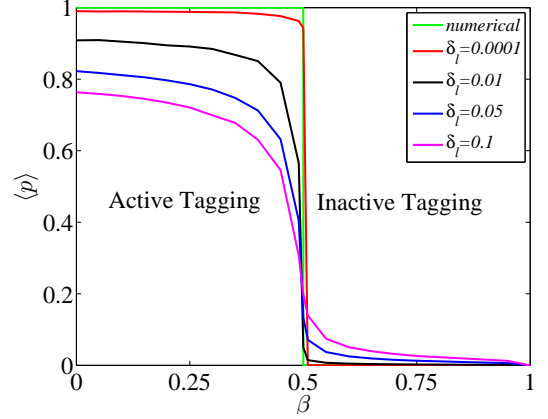


FIG. 4: (Color online) The average tagivity $\langle p \rangle$ as a function of β for various δ_l . The analytical results at $\delta_l \rightarrow 0$ is shown by the green line.

These analytical predictions of $\langle p \rangle$ with $\delta_l = 0$ are compared to simulation results with $\delta_l > 0$. As the results obtained from the two update rules are similar, we present only the results obtained from the linear update. The green line in Fig. 4 shows the analytical stable fixed points of $\langle p \rangle$ with $\delta = 0$. We find that simulations with small δ_l agrees well with the analytical limit, and for $\delta_l > 0$, $\langle p \rangle$ decreases with increasing β as well as increasing δ_l . As we can see, all the simulation results show an abrupt change in $\langle p \rangle$ at $\beta \approx 0.5$, suggesting the existence of a phase transition as predicted by the analytical results. We remark that $\beta = 0.5$ corresponds to the case in Eqs. (5) - (7) where the weight to imitate others equal to that to stay unchanged. These results imply that when users have low confidence, they tend to imitate each other in tagging which leads to a steady state of active tag assignment. However, when users are confident and are stubborn for changes, they stay with their own practice and result in a steady state with inactive tagging. These two behaviors are connected by an abrupt change when confidence increases across $\beta = 0.5$.

To show explicitly how users self-organize to attain the steady state, we start the system at the unstable fixed point and examine how it evolves to the stable fixed point after a slight perturbation. The black line in Fig. 5 corresponds to the average tagivity for the case when confident users (i.e. $\beta < 0.5$) are initialized with zero

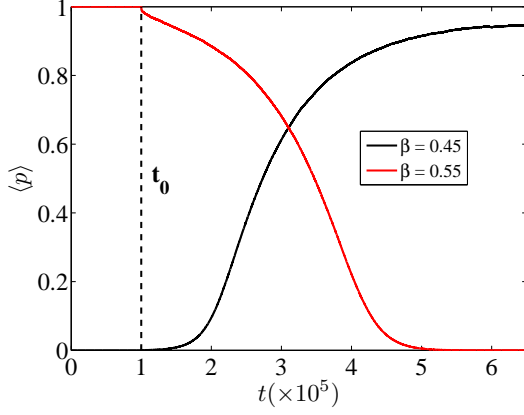


FIG. 5: (Color online) The dynamics of which the self-organization is established. *Black line*: $p_i(0) = 0$ for all users, and at time t_0 one user assigns more than one tags. *Red line*: $p_i(0) = 1$ for all users, and at time t_0 one user assigns only one tag.

tagivity. At time t_0 , one of the users assigns a tag which initiates others to imitate. As we can see, the average tagivity slowly increases after t_0 and saturates at a non-zero steady value, correspond to the self-organization from inactive to active tagging. On the contrary, the red line shows the case when users are initialized with $p_i(0) = 1$ and large confidence (i.e. $\beta > 0.5$). A maximum post length is set to avoid infinite tagging. At time t_0 , one user assigns the minimum number of tags which initiates others to imitate. As we can see, the average tagivity slowly decreases after t_0 and becomes steady at zero, corresponds to the self-organization from active to inactive tagging.

V. EMPIRICAL RESULTS

As it is difficult to define and obtain the tagivity for real users, other well-defined quantities are used for comparison. We compare the distributions of post length obtained from the model with two real datasets: (1) *delicious.com*, a social bookmarking website for saving, sharing and discovering bookmarks associated with tags; (2) *flickr.com*, an image hosting website which encourages users to organize their pictures with tags.

We show in Fig. 6 (a) and (b) the distributions of the post length (as open circles) obtained respectively from *delicious.com* and *flickr.com*. The posts without tags are removed from the statistics. It is interesting to note that the two distributions display similar behaviors: an initial fast decay with post length less than 8, followed by a power-law decay for intermediate post length, and then a high tail. The exponents of the power law decay are 4.1 and 4.3 respectively in *delicious.com* and *flickr.com*, with average post length approximately 2.9 and 3.4. The simulated distributions are plotted in Fig. 6 as *blue* and *red* lines respectively for linear and multiplicative update,

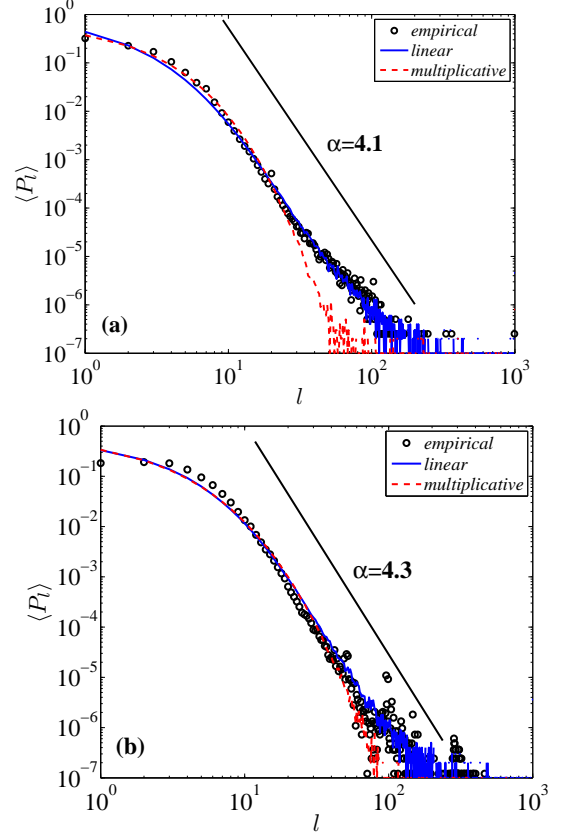


FIG. 6: (Color online) Empirical distribution of the number of tags in each post as compared to simulations. *Circles* represent empirical data, *blue solid lines* and *red dash lines* respectively represent simulation results with linear and multiplicative update. (a) Data from *delicious.com* compared to simulations with parameters $\beta = 0.45$ and $\delta_l = 0.08$ for linear update and $\beta = 0.45$ and $\delta_m = 0.1$ for multiplicative update. (b) Data from *flickr.com* compared to simulations with parameters $\beta = 0.4$, $\delta_l = 0.06$ for linear update and $\beta = 0.4$, $\delta_m = 0.2$ for multiplicative update.

all with $\beta < 0.5$. These results may suggest that real users are of low confidence and tend to imitate each other in tag assignment.

As we can see, the simulation results based on the linear update have better agreement with empirical data than that of the multiplicative update. With the linear update, the high tails of empirical data are also well fitted. According to Fig. 2, the tagivity distribution obtained from the linear update shows a slower decay at large p , as compared to the faster decay in the multiplicative case. The slow decay at large p , i.e. more users are found with large tagging tendency, may explain the high tail in the post length distributions.

VI. CONCLUSIONS AND DISCUSSION

In this paper, we proposed a model to illustrate the self-organization of tagging behaviors in social tagging systems, where individuals imitate each other in tag assignment and eventually result in a self-organized state. With linear update on the tagging tendency, namely *tagivity*, the corresponding steady distribution resembles Gaussian distribution. On the other hand, the steady distribution resembles log-normal distribution when multiplicative update is employed. In addition, we found that when users are of low confidence, they tend to imitate others and the system ends with a steady state of active tagging. By contrast, when users are of high confidence, the system will reach a steady state of inactive tagging. Abrupt changes are observed when user confidence increases and the system changes from one regime to the other, suggesting a phase transition separating the active and inactive tagging. Analyses on convergence time suggest a slow dynamics around the parameter range of phase changes, which provides further evidence for the transition. Finally, the post length distributions of the model are compared to two real datasets obtained from *delicious.com* and *flickr.com*, which show good agreements.

Social tagging systems have been studied with ap-

proaches ranging from graph theory to statistics, which may overlook the interactions and dynamics among individuals. The present model introduced in this paper provides a simple yet interesting description of evolving social tagging systems, which might be generalized to other systems where self-organizations are observed. The proposed model may also shed light on applications (e.g. recommender systems [16, 17]) which combine statistical physics and agent-based models [18] in understanding tagging systems as well as other social systems [19].

Acknowledgment

This work was partially supported by the Program for New Century Excellent Talents in University (NCET-07-0288), the Fundamental Research Funds for the Central Universities and QJectives projects (EU FET-Open Grants 213360 and 231200). ZKZ acknowledges the National Natural Science Foundation of China (Grant nos. 60973069 and 90924011).

References

-
- [1] S.N. Dorogovtsev and J.F.F. Mendes, Phys. Rev. E **63**, 025101 (2001).
 - [2] D. Chowdhury and A. Schadschneider, Phys. Rev. E **59**, R1311 (1999).
 - [3] C. H. Yeung and K. Y. M. Wong, Eur. Phys. J. B **74**, 227 (2010).
 - [4] S. A. Golder and B. A. Huberman, J. Info. Sci. **32**, 198 (2006).
 - [5] C. Cattuto, V. Loreto, and L. Pietronero, Proc. Natl. Acad. Sci. USA **104**, 1461 (2007).
 - [6] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl, Proc. 20th Anniversary Conf. Computer Supported Cooperative Work, pp. 190 (2006).
 - [7] G. Ghoshal, V. Zlatić, G. Caldarelli, and M. E. J. Newman, Phys. Rev. E **79**, 066118 (2009).
 - [8] V. Zlatić, G. Ghoshal, and G. Caldarelli, Phys. Rev. E **80**, 036118 (2009).
 - [9] G. Palla, I. J. Farkas, P. Pollner, I. Deréyi, and T. Vicsek, New J. Phys. **10**, 123026 (2008).
 - [10] Z.-K. Zhang, L. Lü, J.-G. Liu, and T. Zhou, Eur. Phys. J. B **66**, 557 (2008).
 - [11] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto, Proc. Natl. Acad. Sci. USA **106**, 10511 (2009).
 - [12] R. Lambiotte and M. Ausllos, Lect. Notes Comput. Sci. **3993**, 1114 (2006).
 - [13] Z.-K. Zhang and C. Liu, J. Stat. Mech. P10005 (2010).
 - [14] P. Holme and M. E. J. Newman, Phys. Rev. E **74**, 056108 (2006).
 - [15] K. Medvedyeva, P. Holme, P. Minnhagen, and B.J.Kim, Phys.Rev.E **67**, 036118 (2003).
 - [16] Z. -K. Zhang, T. Zhou, and Y. -C. Zhang, Physica A **389**, 179 (2010).
 - [17] Z. -K. Zhang, C. Liu, T. Zhou, and Y. -C. Zhang, EPL **82**, 28002 (2010).
 - [18] E. Bonabeau, Proc. Natl. Acad. Sci. USA **99**, 7280 (2002).
 - [19] C. Castellano, M. Marsili and A. Vespignani, Phys. Rev. Lett. **85**, 3536 (2000).